The Effect of Tandem Substitutions on the Correlation Between Synonymous and Nonsynonymous Rates in Rodents

Nick G. C. Smith and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath B42 7AY, United Kingdom

Manuscript received March 31, 1999
Accepted for publication August 2, 1999

ABSTRACT

Nonsynonymous substitutions in DNA cause amino acid substitutions while synonymous substitutions in DNA leave amino acids unchanged. The cause of the correlation between the substitution rates at nonsynonymous (K_A) and synonymous (K_S) sites in mammals is a contentious issue, and one that impacts on many aspects of molecular evolution. Here we use a large set of orthologous mammalian genes to investigate the causes of the K_A-K_S correlation in rodents. The strength of the K_A-K_S correlation exceeds the neutral theory expectation when substitution rates are estimated using algorithmic methods, but not when substitution rates are estimated by maximum likelihood. Irrespective of this methodological uncertainty the strength of the K_A-K_S correlation appears mostly due to tandem substitutions, an excess of which is generated by substitutional nonindependence. Doublet mutations cannot explain the excess of tandem synonymous-nonsynonymous substitutions, and substitution patterns indicate that selection on silent sites is the likely cause. We find no evidence for selection on codon usage. The nature of the relationship between synonymous divergence and base composition is unclear because we find a significant correlation if we use maximum-likelihood methods but not if we use algorithmic methods. Finally, we find that K_S is reduced at the start of genes, which suggests that selection for RNA structure may affect silent sites in mammalian protein-coding genes.

THE nature of the relationship between nonsynonymous and synonymous substitution rates pertains to many aspects of molecular evolution in mammals (Ina 1996b). A link between the processes of evolution at synonymous and nonsynonymous sites may be due to selection on synonymous sites (see below). Selection on silent sites would affect the selectionist-neutralist debate, for example, providing a potential explanation for the overdispersion of synonymous substitution rates (as shown by Ohta 1995), and would call into question the practice of using silent site comparison to study the evolution of mutation rates (as in McVean and Hurst 1997a).

Several studies have reported a highly significant positive correlation between the synonymous substitution rates (K_S) and the nonsynonymous substitution rates (K_A) of mammalian genes (Wolfe and Sharp 1993; Mouchiroud *et al.* 1995; Makalowski and Boguski 1998b). The K_A-K_S correlation also appears to hold within some mammalian genes (Alvarez-Valin *et al.* 1998).

In this article we investigate a variety of explanations for the intergenic K_A-K_S correlation in mammals, specifically in the comparison between mouse and rat. A number of hypotheses for the K_A-K_S correlation exist (for example see Li 1997). An attractive null hypothesis for the K_A-K_S correlation is the neutral theory explanation, which supposes that genes differ in mutation rates, that all synonymous changes are neutral, and that a variable proportion of nonsynonymous changes are neutral (Ohta and Ina 1995). This null hypothesis can explain the existence of a K_A-K_S correlation in mammals but appears unable to explain why the correlation is so strong (Ohta and Ina 1995). We confirm this result using an improved data set and algorithmic rate estimation methods, but we also show that the K_A-K_S correlation is consistent with the neutral prediction if one uses maximum likelihood (ML) to estimate substitution rates (see results). Thus methodological bias may have led to previous overestimates of the strength of the K_A-K_S correlation.

Despite the fact that the strength of the K_A-K_S correlation may be consistent with silent site neutrality, patterns of substitutions indicate that selection may well be acting on silent sites. In particular, the strength of the K_A-K_S correlation appears in large part to be due to an excess of tandem substitutions caused by substitutional nonindependence.

Synergy between synonymous and nonsynonymous substitutions, such that one type of substitution increases the likelihood of the other, would increase the K_A-K_S correlation. Such substitutional nonindependence could be the result of either selection or muta-

Corresponding author: Nick G. C. Smith, School of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom.
E-mail: n.g.c.smith@sussex.ac.uk

tion. Purifying selection might act on both nonsynonymous and synonymous sites (Ina 1996a), or nonsynonymous substitutions might cause positive selection on subsequent synonymous substitutions (Lipman and Wilbur 1985). Alternatively, a single mutational event might affect both synonymous and nonsynonymous sites simultaneously as with doublet mutations (Wolfe and Sharp 1993). (Note on terminology: we use "doublet" to refer to a supposed mutational event affecting adjacent bases and "tandem" to apply to observed adjacent substitutions.)

It is also possible to envisage a hybrid selection-mutation model in which a correlation between the mutation rate and nonsynonymous constraints causes an increase in the K_A-K_S correlation (Ina 1996a). Such a hybrid explanation is supported by theoretical (Kondrashov 1995) and empirical (McVean and Hurst 1997b; Smith and Hurst 1999) studies of the evolution of mutation rates.

MATERIALS AND METHODS

Selection of protein coding sequences: A list of 470 genes in mouse, rat, and human, with orthology confirmed using HOVERGEN 19 (Duret *et al.* 1994), was obtained from Makalowski and Boguski (1998a). Only genes with complete protein-coding sequence available in a single GenBank/EMBL record were used, leaving 432 three-species comparisons.

Preparation of alignments: Alignments were performed using the GCG (1994) and EGCG (Rice 1997) packages at HGMP (http://www.hgmp.mrc.ac.uk/). FETCH was used to extract sequences from databases, and GENETRANS was used to extract and combine exons automatically. Protein alignments were performed using CLUSTALW (Thompson *et al.* 1994). Then the DNA alignments were recreated from the protein alignments and the original DNA sequences using the program MRTRANS (written by W. Pearson and available at HGMP).

ML analysis: The ML package PAML (Yang 1997) was used to reconstruct ancestral sequences and to estimate substitution rates. We used the program BASEML to reconstruct ancestral sites, with the gene tree defined as [(mouse, rat), human], with no rate variation between sites, and with the REV model of evolution. Ancestral sequence reconstruction was carried out by ML, rather then parsimony, for two reasons: ML allows the reconstruction of all sites, and parsimony is biased when base composition is skewed (Eyre-Walker 1998).

The program CODEML, under a codon-based model of evolution (Goldman and Yang 1994), was used to estimate K_A and K_S. Using PAML version 2.0 the following parameter settings were used: seqtype = 1, codon-based model; runmode = −2, estimate K_A and K_S rates; CodonFreq = 3, codon frequencies used as free parameters; additionally, no rate variation was allowed.

Algorithmic rate estimation: Substitution rates were also estimated from sequence alignments using algorithmic methods developed by Moriyama and Powell (1997). Tamura's (1992) multiple hits correction method was used in conjunction with Li's (1993) method to calculate K_A and K_S. The substitution rates at fourfold synonymous sites, K_4, were also estimated using the algorithmic method of Tamura and Nei (1993). Estimates of K_4 are expected to be more reliable than estimates of K_S, which have to combine the rates of sites of different degeneracies.

With regard to the differences between the algorithmic and PAML rate estimation methods, the algorithmic methods gave similar results to PAML using CodonFreq = 1, codon frequencies calculated from average nucleotide frequencies. But with PAML using CodonFreq = 2, codon frequencies calculated from average nucleotide frequencies at the three codon positions, and PAML using CodonFreq = 3, codon frequencies as free parameters, the PAML and algorithmic estimates differed with regard to the strengths of the K_A-K_S and the K_S-composition correlations (data not shown, but see results for a comparison of the algorithmic estimates and PAML estimates using CodonFreq = 3).

Measurement of substitutional nonindependence: To analyze lineage-specific substitution patterns, we used mouse, rat, and human orthologs to reconstruct ancestral sequences (see above) and compared present-day sequences to their most recent ancestral node. The mouse and rat lineage-specific substitution patterns were combined.

The measurement of substitution patterns proceeded as follows. Substitutions between two sequences were designated as either fully synonymous (syn) or fully nonsynonymous (nonsyn) or mixed (part syn and part nonsyn), following the method of Li *et al.* (1985). All substitutions within 100 bp of every other substitution were investigated, and the totals of all substitution pairs a certain distance apart were noted (if one or both of the substitutions was mixed the necessary weightings were applied, and indels were ignored). Three classes of substitution pairs were investigated: syn-syn, syn-nonsyn, and nonsyn-nonsyn.

Simulated substitution sequences were generated under the assumption of independent substitutions. Simulated sequences were the same length as the real sequences and were generated according to the codon position-specific synonymous and nonsynonymous substitution rates of the real sequences so that the substitution rates of the simulations were the same as those of the real sequences. The same substitution pattern analysis was performed on the simulated sequences as on the real sequences. For each sequence considered the substitution patterns of the real sequence were compared against those of 500 simulated sequences.

Statistics describing the difference between the real and simulated substitution patterns were calculated for all three substitution pair classes. The greater the difference between the real and simulated substitution patterns the greater the nonindependence between real substitutions, and thus we term our statistic substitutional nonindependence (SNI). The numbers of real cases (r) were summed for all N sequences, and for each simulation run the numbers of simulated occurrences (s) were summed for all sequences. SNI is given by the number of simulation runs for which the real total was greater than the simulated total, so for 500 simulations per sequence we have the formula

$$\mathrm{SNI} = \sum_{j=1}^{500}\left(\sum_{i=1}^{N} r_i > \sum_{i=1}^{N} s_{ij}\right).$$

Under the null assumption of no difference between real and simulated substitution patterns, the expected value of SNI is 250. Using the normal distribution as an approximation to the binomial, we find the one-tailed 95% upper confidence limit to be 268. If we apply the Bonferroni correction for considering 100 different substitution pair distances (as described on page 240 in Sokal and Rohlf 1995), the upper confidence limit is 286.

As an aid to visualization of substitutional nonindependence, we have also provided plots of substitution pair class separation against the statistic real over simulated (ROS),

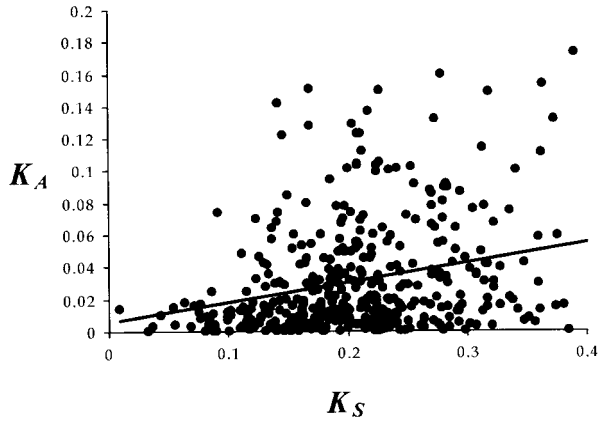

Figure 1.—K_A plotted against K_S for 432 mouse-rat genes. Substitution rates were estimated using maximum-likelihood methods. The linear regression line is shown.

which increases from unity upward as substitutional nonindependence increases, and which is defined as

$$\mathrm{ROS} = \frac{(1/N)\Sigma_{i=1}^{N} r_i}{(1/500N)\Sigma_{j=1}^{500}\Sigma_{i=1}^{N} s_{ij}}.$$

RESULTS

The K_A-K_S correlation is consistent with neutral theory: Using ML rate estimation and a large set of orthologous mouse-rat genes (see materials and methods and Figure 1), we estimated the K_A-K_S correlation coefficient by rank correlation followed by the *z*-transformation (Sokal and Rohlf 1995). Contrary to the suggestion that a significant K_A-K_S correlation results from the inclusion of paralogs (Hughes and Yeager 1997), and in agreement with previous findings (Mouchiroud *et al.* 1995; Makalowski and Boguski 1998b), we find a highly significant positive correlation between K_A and K_S ($P < 0.0001$) using both algorithmic and ML rate estimation methods.

In addition we calculated the K_A-K_S correlation coefficients predicted by the neutral theory explanation as given by Ohta and Ina (1995). In agreement with their results, we find that the neutral theory is unable to explain the strength of the observed K_A-K_S correlation if algorithmic rate estimation methods are used, with the observed correlation coefficient R greater than the expected correlation coefficient ρ ($R = 0.411$ and $\rho = 0.270$). Statistical testing is difficult because the variance of ρ is not theoretically tractable (Ohta and Ina 1995), but simulations have shown that findings of $R \gg \rho$ can be explained by pure chance (Ina 1996a).

However, in contrast to the results of Ohta and Ina (1995), we find that the neutral theory is consistent with the strength of the observed K_A-K_S correlation if ML rate estimation methods are used, with R less than but similar to ρ ($R = 0.275$ and $\rho = 0.343$). The evolutionary model specified in PAML is more general than that of the algorithmic method, which might lead one to conclude that the PAML rate estimates are probably more reliable. However, the PAML rate estimates should not be considered perfect: standard errors, required to predict ρ, are estimated using the normal approximation to the likelihood curve; and the model of evolution makes no allowance for rate variation between sites. There is also the question of whether pairwise sequence comparisons provide enough data for the ML approach to provide unbiased estimates. We conclude that it is unclear which of the algorithmic or ML approaches is more reliable and thus can only note the methodological sensitivity of the strength of the K_A-K_S correlation relative to the neutral theory prediction.

The importance of tandem substitutions: The influence of tandem substitutions was investigated using ML rate estimation (similar results were obtained using algorithmic methods). If tandem substitutions were ignored, the expected correlation coefficient considerably exceeded the observed correlation coefficient ($R = 0.046$ and $\rho = 0.349$); thus tandem substitutions appear to make a large contribution to the strength of the K_A-K_S correlation. Upon removal of tandem substitutions the ratio of the expected correlation coefficient to the observed correlation coefficient changes from 1.25 to 7.59, a sixfold increase.

If only those genes with no tandem substitutions were considered ($N = 67$), the K_A-K_S correlation was zero, considerably below the neutral expectation ($R = 0$ and $\rho = 0.344$). This result suggests that the K_A-K_S correlation is generated almost exclusively by tandem substitutions, although this interpretation should be treated with caution as the genes with no tandem substitutions were atypically short and slowly evolving (data not shown).

Substitutional nonindependence mainly affects adjacent bases: The K_A-K_S correlation is strengthened if there is substitutional nonindependence between synonymous and nonsynonymous sites (see Introduction). The effect of tandem substitutions on the K_A-K_S correlation implies nonindependence between adjacent substitutions; but does substitutional nonindependence occur at other distances? We measured the nonindependence between syn-nonsyn pairs of substitutions at all pair separation distances from 1 to 100 bases (see materials and methods). If all substitutions are considered, then substitutional nonindependence appears to operate at a variety of distances: 80 of the 100 syn-nonsyn pairs have highly significant SNI values ($P < 0.05$ with Bonferroni correction). The ROS plot (Figure 2) shows high levels of substitutional nonindependence for the syn-nonsyn pairs, with ROS values tending to decrease as the distance between the two substitutions increases (note that tandem syn-nonsyn substitutions give the highest ROS value).

To check whether substitutional nonindependence really exists beyond effects between adjacent bases, we investigated the effect of the removal of tandem substitu-

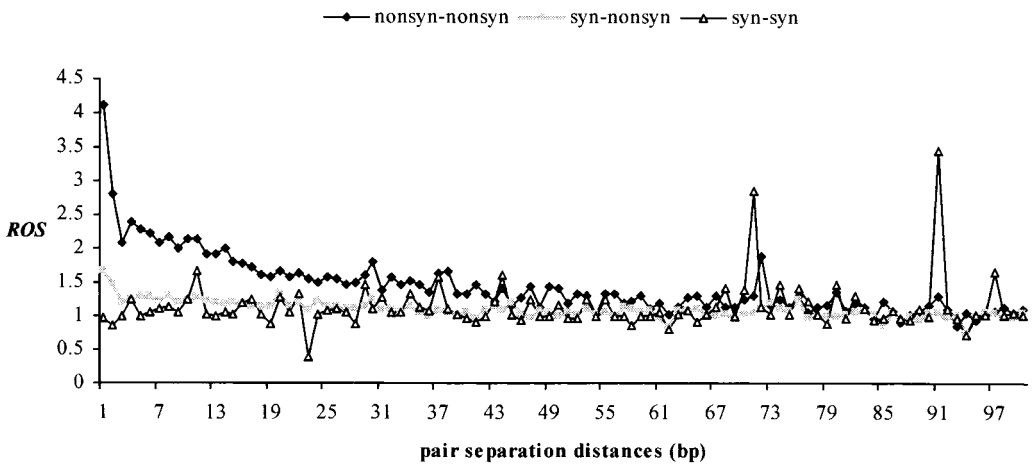

Figure 2.—The ROS plots of substitutional nonindependence for the rodent lineages. Values are given for all three classes of substitution and for all pair separation distances from 1 to 100 bp (see materials and methods).

tions on substitution patterns. The resultant change in patterns of substitutional nonindependence is striking (compare Figures 2 and 3). Not a single syn-nonsyn pair yielded a significantly high SNI value ($P > 0.05$ without Bonferroni correction). These results imply that whatever process (selection or mutation) is responsible for the nonindependence of syn-nonsyn substitutions, then that process is mainly acting on adjacent bases and causing an excess of tandem substitutions.

The excess of tandem substitutions is not due to doublet mutations: If mutational processes are sufficient to explain the excess of tandem substitutions without recourse to selection, then synonymous changes are neutral and doublet mutations are responsible for the excess of tandem substitutions. From these assumptions we can predict an excess of neighboring syn-syn pairs. But the SNI value for neighboring syn-syn pairs is 143, which is lower than the null expectation of 250. This means that either doublet mutations do not occur or that synonymous doublet mutations are subject to purifying selection. Either way, we can conclude that mutation alone is unable to explain the excess of tandem substitutions. Hence by elimination we are left with a selective explanation for the excess of tandem substitutions.

Selection on silent sites is demonstrated by patterns of substitutional nonindependence: We have shown that synonymous-nonsynonymous substitutional nonindependence does not appear to exist beyond the interactions of adjacent bases. Given that we have also provided evidence against doublet mutations, we have no reason to believe in any form of mutational nonindependence. If we make the assumption that mutation does not differentiate between synonymous and nonsynonymous sites, then we can conclude that any differences in substitutional nonindependence between the three classes of substitution pairs (syn-syn, syn-nonsyn, and nonsyn-nonsyn) must be due to selection.

The different types of substitution pairs do indeed show significantly different levels of substitutional nonindependence. For each class of pairs, 100 different measures of SNI were obtained, corresponding to all the pair separation distances from 1 to 100 bp. Out of a possible maximum of 100, 96 of the nonsyn-nonsyn pair classes, 80 of the syn-nonsyn pair classes, and 69 of the syn-syn pair classes have highly significant SNI values ($P < 0.05$ with Bonferroni correction). The ROS plots (Figure 2) show the same pattern of substitutional nonindependence decreasing in the order of nonsyn-nonsyn, syn-nonsyn, and syn-syn. Both the nonsyn-nonsyn

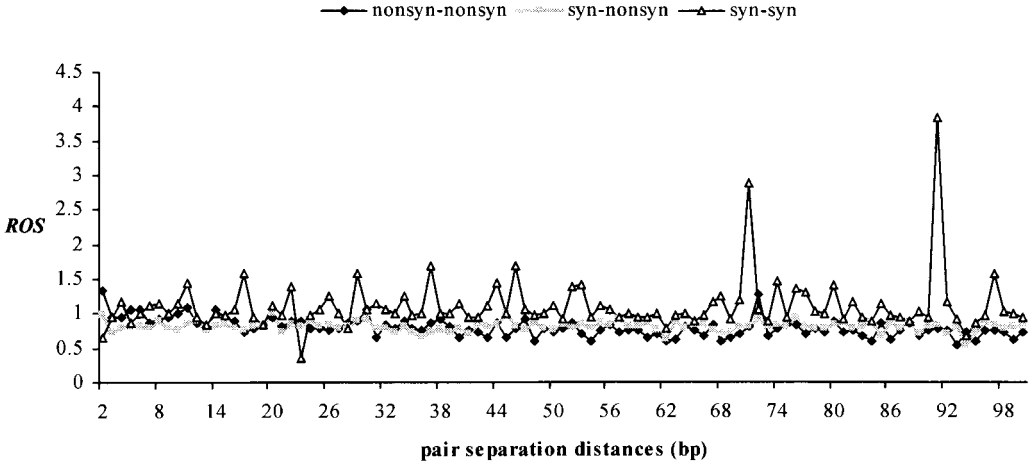

Figure 3.—The ROS plots of substitutional nonindependence for the rodent lineages after removal of tandem substitutions. Compare with Figure 2.

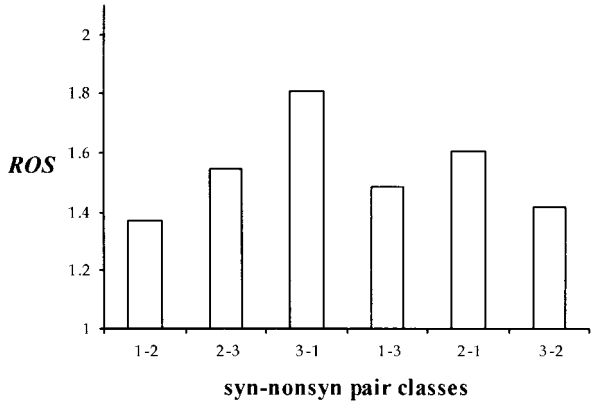

Figure 4.—The ROS measures of substitutional nonindependence for six classes of syn-nonsyn pairs as defined by the codon positions of the substitutions. If selection acts on codon usage, then those pairs contained within a single codon (1-2, 2-3, and 1-3) should have higher ROS values than the other pairs (3-1, 2-1, and 3-2).

(Mann-Whitney *U*-test, $P < 0.0001$) and syn-nonsyn (Mann-Whitney *U*-test, $P = 0.04$) pair classes show significantly greater substitutional nonindependence than the syn-syn pair class.

These results do not appear to be the result of unreliable ancestral sequence reconstruction, because qualitatively identical results are obtained from the mouse-rat interspecies comparison as from the lineage-specific comparisons (data not shown). Therefore selection appears to be operating on silent sites, though we accept that our conclusion is based on an assumption concerning the nature of the mutational process. We now attempt to discern the precise nature of the selection on silent sites.

Selection for major codon usage: If selection acts to favor major codon usage (Akashi and Eyre-Walker 1998), then substitutional nonindependence should be greater for pairs of substitutions within codons than pairs of substitutions between codons.

The syn-nonsyn substitution pairs at distances of 1 and 2 bp were both divided into three classes according to the codon positions of the substitutions. The pairs 1 bp apart were classified as 1-2, 2-3, and 3-1. Both 1-2 and 2-3 represent a pair of substitutions within a codon, while 3-1 invokes substitutions in adjacent codons. Similarly, the pairs 2 bp apart were classified as 1-3, 2-1, and 3-2. In this case only 1-3 comprises substitutions within a codon, while both 2-1 and 3-2 involve substitutions in adjacent codons.

All six substitution pair classes show highly significant SNI values ($P < 0.05$ with Bonferroni correction), and thus the SNI data are equivocal on the issue of selection for codon usage. The ROS data are contrary to predictions based on selection for codon usage: ROS is greater in the 3-1 class than in the 1-2 and 2-3 classes, and ROS in the 1-3 class is intermediate between that in the 2-1 and 3-2 classes (see Figure 4). Our finding of no evidence in favor of selection for major codon usage in mammals supports previous studies (Eyre-Walker 1991; Smith and Hurst 1999).

Selection for base composition: The relationships between synonymous substitution rates and a number of compositional characters were examined to test predictions of specific selective pressures. Significant correlations would be consistent with selection acting directly on base composition or a link between selection and other characters that correlate with composition (such as recombination; Eyre-Walker 1993). However, this test is not capable of providing strong evidence in favor of selection, because K_S-composition correlations could be the result of mutation rather than selection.

As with the K_A-K_S correlation, the alternative methods of rate estimation yield different results. With the algorithmic method K_S does not correlate strongly with either *GC4* (G plus C content at fourfold degenerate sites; $R = 0.008$), *A4* ($R = -0.03$), *C4* ($R = -0.025$), *G4* ($R = 0.071$), or *T4* ($R = -0.007$). Using the more reliable algorithmic measure of K_4 we also find no correlation between synonymous divergence and base composition (*GC4* and K_4; $R = 0.002$; see Figure 5). However, with PAML we find significant correlations ($P < 0.0001$) for all compositional parameters: *GC4* ($R = 0.258$; see Figure 5), *A4* ($R = -0.264$), *C4* ($R = 0.187$), *G4* ($R = 0.247$), and *T4* ($R = -0.206$).

These differences between the methods are all the more surprising when one considers that, as one would expect, the alternative measures of synonymous divergence are highly significantly correlated ($R \sim 0.9$). Given that we are unable to choose between algorithmic and ML methods (see above), these data are equivocal on the issue of selection on silent sites (for evidence of selection on the base composition of mammalian silent sites, see Eyre-Walker 1999). However, our results are pertinent to the debate as to whether there is a relationship between K_S and base composition. The existence of a significant correlation was originally suggested by Wolfe *et al.* (1989) on the basis of a fairly small sample. Bernardi *et al.* (1997) subsequently showed that the inverted *V* distribution obtained by Wolfe *et al.* (1989) was at least partially due to rate estimate biases (see Pesole *et al.* 1995). However, our ML results suggest a linear relationship between *GC4* and K_S (see Figure 5), which cannot be so easily explained by methodological biases.

Selection for RNA structure: Selection on RNA structure has been proposed as an explanation for the reduced K_S at the start of protein-coding enterobacterial genes, with an open structure thought to favor ribosome binding (Eyre-Walker and Bulmer 1993). We have found a similar pattern in our set of mammalian genes (see Figure 6). For all 354 genes with mouse-rat alignments longer than 600 bp, K_S was estimated using algorithmic methods for five regions of the gene: the whole gene and the first four nonoverlapping sections of 50

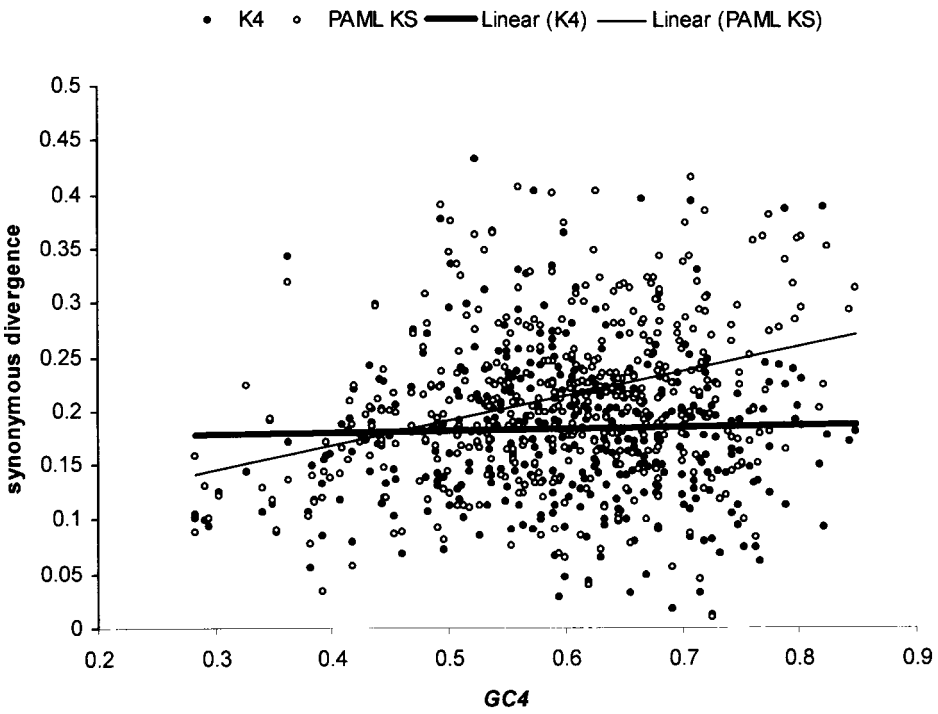

Figure 5.—Base composition at fourfold degenerate sites (*GC4*) plotted against synonymous divergence for 432 mouse-rat genes. Two measures of synonymous divergence are shown: PAML K_S is a maximum-likelihood estimate while K_4 is an algorithmic estimate (see materials and methods). The linear regression lines show a significant relationship between PAML K_S and *GC4* but not between K_4 and *GC4*.

codons. The first 50 codons at the start of the gene have a significantly low K_S in comparison to both the whole gene (Mann-Whitney *U*-test, $P < 0.0001$) and three subsequent 50-codon blocks (Mann-Whitney *U*-tests, $P = 0.0019$, $P = 0.0081$, $P = 0.0047$). These findings provide us with suggestive, although by no means conclusive, evidence that silent sites in mammals are affected by selection.

It is thought that longer mRNAs have a lower density of longer stem loops, and so selection on RNA structure is predicted to decrease with increasing gene length (Comeron and Aguade 1996). We find no correlation between gene length and K_S for either rate estimation method, though we note that this appears to be a weak test of selection.

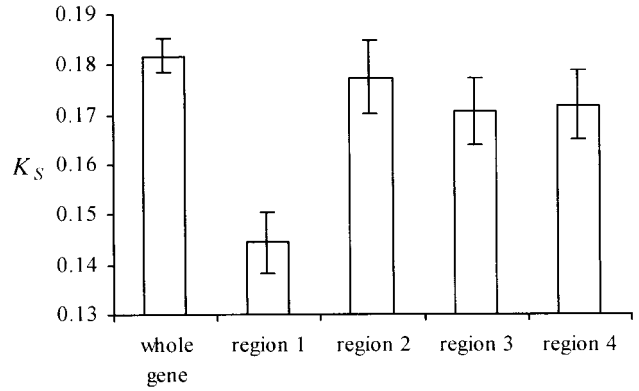

Figure 6.—The first 50 codons of a gene (region 1) have a low K_S relative to the whole gene and the three subsequent 50-codon nonoverlapping sections (regions 2 to 4). The error bars indicate the SEs of the means. Rates were estimated using algorithmic methods.

DISCUSSION

With a ML approach to rate estimation, the rodent K_A-K_S correlation coefficient is consistent with the neutral theory, but using an algorithmic approach the correlation is stronger than expected. Despite such methodological uncertainty we have found strong evidence to suggest that the excess of tandem substitutions generated by substitutional nonindependence contributes to the strength of the rodent K_A-K_S correlation coefficient. The removal of tandem substitutions reduces the K_A-K_S correlation coefficient by a factor of six, and there exists no K_A-K_S correlation for those genes that do not contain tandem substitutions. Substitutional nonindependence between adjacent bases, the process that generates the excess of tandem substitutions, appears to be the dominant form of substitutional nonindependence.

What causes the excess of tandem substitutions that contribute to the K_A-K_S correlation? Is it selection or mutation? We demonstrate that the mutational explanation fails due to a lack of evidence for doublet mutations, which means that selection must be responsible for the excess of synonymous-nonsynonymous tandem substitutions. Our analysis of the substitution patterns of the different pair classes also supports the notion of silent site selection, and encourages us to investigate the form of selection acting on silent sites. It might be argued that our finding of substitutional nonindependence caused by selection is inconsistent with our finding using ML methods that the K_A-K_S correlation is consistent with neutrality, but the neutral prediction should remain reasonably accurate as long as the proportion of silent sites affected by selection is low. Although tandem substitutions are contributing greatly to the K_A-K_S correla-

tion, selection may generate a relatively small excess of tandems above those predicted on the basis of neutrality.

By examining substitution patterns we have provided evidence against selection acting on codon usage. We have found that the existence of correlations between K_S and base composition depends on rate estimation methodology and offers no clue as to whether selection via base composition acts on silent sites. There is no correlation between K_S and gene length, but selection on RNA structure is consistent with our finding that K_S is reduced at the start of mammalian genes. Although further work is clearly required to examine this supposition, we suggest that selection on RNA structure is a possible explanation for the strong syn-syn substitutional nonindependence at distances of 71 and 91 bp (see Figure 2).

What are the implications of our results with respect to mammalian molecular evolution? We have found three reasons to believe that silent sites in mammals are subject to selection: (i) mutation cannot explain the excess of syn-nonsyn tandem substitutions, therefore selection is responsible by elimination; (ii) a comparison of the levels of substitutional nonindependence of the syn-syn, syn-nonsyn, and nonsyn-nonsyn classes of substitution pairs appears to indicate the effects of selection; and (iii) low K_S at the start of genes is consistent with selection on RNA structure. Although arguments (ii) and (iii) are by no means certain, we consider reason (i) to provide strong evidence for silent site selection.

Selection on silent sites can explain the overdispersion of silent sites in mammals (as in Ohta 1995). But does silent site selection necessarily invalidate those studies of the evolution of the mutation rate in mammals, which assume that silent sites are neutral and hence that K_S can be used as an unbiased estimator of the mutation rate (as in McVean and Hurst 1997a)? Although we have found evidence of selection on silent sites we still believe that K_S provides the best available estimate of the mutation rate. First, K_S values before and after the removal of tandem substitutions are highly significantly correlated (using PAML, $R = 0.927$ and $P < 0.00001$). Second, tests of adaptive mutation rates hold both before and after the removal of tandem substitutions (Smith and Hurst 1999). Third, there is a practical argument in favor of using K_S, which is that the alternative way to estimate mutation rates is to use noncoding DNA sequence data, the alignment of which is problematic (Smith and Hurst 1998).

The authors thank Ziheng Yang, Yasuo Ina, Adam Eyre-Walker, Paul Higgs, and Jonathan Slack. L.D.H. is funded by the Royal Society.

LITERATURE CITED

Akashi, H., and A. Eyre-Walker, 1998 Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8:** 688–693.

Alvarez-Valin, F., K. Jabbari and G. Bernardi, 1998 Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. **46:** 37–44.

Bernardi, G., D. Mouchiroud and C. Gautier, 1997 Isochores and synonymous substitutions in mammalian genes, pp. 137–168 in *DNA and Protein Sequence Analysis*, edited by M. J. Bishop and C. J. Rawlings. Oxford University Press, Oxford.

Comeron, J. M., and M. Aguade, 1996 Synonymous substitutions in the Xdh gene of *Drosophila*—heterogeneous distribution along the coding region. Genetics **144:** 1053–1062.

Duret, L., D. Mouchiroud and M. Gouy, 1994 Hovergen—a database of homologous vertebrate genes. Nucleic Acids Res. **22:** 2360–2365.

Eyre-Walker, A., 1991 An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33:** 442–449.

Eyre-Walker, A., 1993 Recombination and mammalian genome evolution. Proc. R. Soc. Lond. Ser. B **252:** 237–243.

Eyre-Walker, A., 1998 Problems with parsimony in sequences of biased base composition. J. Mol. Evol. **47:** 686–690.

Eyre-Walker, A., 1999 Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152:** 675–683.

Eyre-Walker, A., and M. Bulmer, 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21:** 4599–4603.

GCG, 1994 *Program Manual for the Wisconsin Package, Version 8*. Genetics Computer Group, Madison, WI.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Hughes, A. L., and M. Yeager, 1997 Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. **45:** 125–130.

Ina, Y., 1996a Correlation between synonymous and nonsynonymous substitutions and variation in synonymous substitution numbers, pp. 105–113 in *Current Topics on Molecular Evolution*, edited by M. Nei and N. Takahata. Institute of Molecular Evolutionary Genetics, Penn State University, University Park, PA and The Graduate University for Advanced Studies, Hayama, Japan.

Ina, Y., 1996b Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. J. Genet. **75:** 91–115.

Kondrashov, A. S., 1995 Modifiers of mutation-selection balance: general-approach and the evolution of mutation-rates. Genet. Res. **66:** 53–69.

Li, W. H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

Li, W. H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Li, W. H., C. I. Wu and C. C. Luo, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

Lipman, D. J., and W. J. Wilbur, 1985 Interaction of silent and replacement changes in eukaryotic coding sequences. J. Mol. Evol. **21:** 161–167.

Makalowski, W., and M. S. Boguski, 1998a Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA **95:** 9407–9412.

Makalowski, W., and M. S. Boguski, 1998b Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. J. Mol. Evol. **47:** 119–121.

McVean, G. T., and L. D. Hurst, 1997a Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature **386:** 388–392.

McVean, G. T., and L. D. Hurst, 1997b Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. Proc. R. Soc. Lond. Ser. B **264:** 739–746.

Moriyama, E. N., and J. R. Powell, 1997 Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J. Mol. Evol. **45:** 378–391.

Mouchiroud, D., C. Gautier and G. Bernardi, 1995 Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. **40:** 107–113.

Ohta, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. **40:** 56–63.

Ohta, T., and Y. Ina, 1995 Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J. Mol. Evol. **41:** 717–720.

Pesole, G., G. Dellisanti, G. Preparata and C. Saccone, 1995 The importance of base composition in the correct assessment of genetic distance. J. Mol. Evol. **41:** 1124–1127.

Rice, P., 1997 *Program Manual for the EGCG Package.* The Sanger Centre, Hinxton Hall, Cambridge, CB10 1RQ, England.

Smith, N. G. C., and L. D. Hurst, 1998 Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. **47:** 493–500.

Smith, N. G. C., and L. D. Hurst, 1999 The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics **152:** 661–673.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry.* W. H. Freeman and Company, New York.

Tamura, K., 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Mol. Biol. Evol. **10:** 512–526.

Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 ClustalW—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Wolfe, K. H., and P. M. Sharp, 1993 Mammalian gene evolution—nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37:** 441–456.

Wolfe, K. H., P. M. Sharp and W. H. Li, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

Communicating editor: G. B. Golding